Exp Brain Res (2012) 216:457–462
DOI 10.1007/s00221-011-2948-9

RESEARCH ARTICLE

Perceived size change induced by audiovisual temporal delays

Philip Jaekl · Salvador Soto-Faraco · Laurence R. Harris

Received: 15 September 2011 / Accepted: 8 November 2011 / Published online: 22 November 2011

Abstract The retinal image of an object does not contain information about its actual size. Size must instead be inferred from extraretinal cues for which distance information makes an essential contribution. Asynchronies in the arrival time across visual and auditory sensory components of an audiovisual event can reliably cue its distance, although this cue has been largely neglected in vision research. Here we demonstrate that audio-visual asynchronies can produce a shift in the apparent size of an object and attribute this shift to a change in perceived distance. In the present study participants were asked to match the perceived size of a test circle paired with an asynchronous sound to a variable-size probe circle paired with a simultaneous sound. The perceived size of the circle increased when the sound followed its onset with delays up to around 100 ms. For longer sound delays and sound leads, no effect was seen. We attribute this selective modulation in perceived visual size to audiovisual timing influences on the intrinsic relationship between size and distance. This previously unsuspected cue to distance reveals a surprisingly interactive system using multisensory information for size/distance perception.

P. Jaekl (✉) · S. Soto-Faraco
Departament de Tecnologies de la Informació i les Comunicacions, Grup de Recerca Neurociencia Cognitiva
Universitat Pompeu Fabra, Barcelona, Spain
e-mail: phil.jaekl@upf.edu

S. Soto-Faraco
Institució Catalana de Recerca i Estudis Avancats (ICREA), Barcelona, Spain

L. R. Harris
Centre for Vision Research, York University, Toronto, Canada

Keywords Audiovisual integration · Size/distance invariance · Size perception · Distance perception

Introduction

Information about an object's size (either absolute or relative) cannot be reliably extracted from its retinal image. This is obvious for very distant objects such as the moon. The perception of the moon's absolute size is completely inaccurate and judgments about its relative size can change dramatically as demonstrated by the moon illusion, usually attributed to changes in the moon's perceived distance (see Hershenson 1989; Ross and Plug 2002 for reviews). The moon illusion, and a host of similar size illusions such as the Ames Room, emphasize how perceived linear size is obtained from extraretinal information (see Gregory 1997). Here we address whether audiovisual multisensory processing can contribute to size/distance estimation.

Sound can exert many intriguing effects on visual perception resulting in changes in perceived brightness, duration, and visibility (see Shams and Kim 2010 for a recent review). More relevantly, these effects include the selective integration of auditory and visual motion cues for looming stimuli (Cappe et al. 2009), an auditory aftereffect in which a sound appears to change in loudness following adaptation to visual motion in depth (Kitagawa and Ichihara 2002; Valjamae and Soto-Faraco 2008) and even a cross-modal synesthetic association between auditory frequency and visual size (Gallace and Spence 2006; Parise and Spence 2009). Audiovisual interactions of this type support the potential interplay between distance cues across modalities, but no evidence directly supports auditory influences on the perceived size of a visual object.

Any sensory information that affects the perceived distance of an object is likely to induce a change in its perceived

size since retinal image size is proportional both to the linear size of the object being imaged and its distance from the observer (Schlosberg 1950; Gilinsky 1951; Kilpatrick and Ittelson 1953; Gilinsky 1955; Epstein et al. 1961; Kaufman et al. 2006; Kaufman et al. 2007). Previous investigations of distance cues have tended to regard distance estimation as a unisensory visual or auditory computation, providing separate, non-interacting estimates. However, many events provide both auditory and visual signals. Moreover, light and sound travel through air at very different velocities so that visual cues about a audiovisual event arrive at an observer virtually instantaneously while sound arrives after a delay that varies linearly with distance. This cross-modal asynchrony due to physical distance has been considered a source of variability requiring compensation for accurate perception (Spence and Squire 2003; Kopinska and Harris 2004; King 2005). Yet, auditory-visual asynchrony contains precise, linear information that can potentially be used to estimate the absolute distance to an event.

We therefore investigated whether auditory-visual asynchrony could induce a change in the perceived size of a visual object. Participants compared the size of a test circle paired with delayed sound with a variable-sized probe circle paired with simultaneous sound. We hypothesized that, with angular size constant, sound delay would induce an increase in perceived size congruent with the perception of the circle appearing more distant. In a subsequent experiment participants were tested on a wider range of asynchronies including sound both leading and lagging the onset of the test circle. This second experiment aimed to probe the temporal window of the effect and to determine if a sound leading the visual stimulus induced a decrease in perceived size, consistent with the perception of it being closer.

Psychophysical investigations of audiovisual interaction have demonstrated various sound-induced enhancements of visual properties. Such enhancements include increases in detectability, perceived brightness, temporal resolution, duration, motion detection, and localization of visual attention (see Shams and Kim 2010 for a review). Importantly, such effects are greatest when the auditory and visual components of the event are approximately synchronous. In the current experiment we present clearly visible stimuli and look for size changes dependent not on synchrony but on asynchrony in arrival times.

Method

Participants

Fifteen participants (9 female, mean age 24 years, range = 20–31 years) completed experiment one in which audiovisual stimuli consisted of visual targets presented only with delayed sound onsets and a baseline, synchronous condition. A second experiment was completed by nine participants (5 female, mean age = 27 years, range = 19–55 years) which included sound-leading SOAs. All participants gave informed consent and were paid €5. They had normal hearing and normal or corrected-to-normal vision. All procedures were approved by the local ethical committee at UPF.

Apparatus

All experiments were carried out in a completely dark room. Visual stimuli were created using a PC running Matlab version 2010a in conjunction with the Psychophysics Toolbox extensions version 3 (Brainard 1997) and displayed on a 21″ Sony Trinitron flat screen CRT monitor (refresh rate 100 Hz, resolution 800 × 600 pixels). Two loudspeakers (Altec Lansing VS2420) presented sound simultaneously from each side of the monitor (±2.5°). Auditory stimuli were processed through a Creative Sound Blaster Audigy 2 ZS sound card in conjunction with the Psychtoolbox PsychPortAudio command library. Accurate timing calibration between auditory and visual stimulus presentation was confirmed using a Black Box Toolkit (Black Box Toolkit Ltd., England) which compared signals from a photosensitive diode and a digital microphone.

Stimuli

Visual stimuli were filled circles of luminance 1 cd/m^2 presented for 200 ms on a blank background (0.1 cd/m^2). The test circle had a diameter of 1.3° with the probe circle varying around this value (see below). Auditory stimuli were 9.6 kHz, square-wave tones (intensity 70 dB(A)) measured at the participant's ear level. For the first experiment, the delay between the test circle and the tone varied from 0 to +120 ms in 20 ms increments. In the second experiment, the delays were ±128, ±64, ±32, ±16, and 0 ms (negative corresponding to sound leading the visual stimulus). Apart from the difference in delays, the two experiments were identical.

Procedure

To prevent participants knowing the distance of the display, they were taken into the experimental room blindfolded and seated with the help of an assistant. Participants dark-adapted for approximately 5 to 10 mins while they were given instructions. Viewing was monocular through the dominant eye. Participants were told they could keep the other eye open under a patch and were seated 5 m from the screen with their head stabilized on a chinrest.

Each trial began with a 1.5 s presentation of a fixation crosshair (0.5°) followed by 1 s of blank screen.

Participants then saw a test and a probe stimulus for 200 ms each, presented in random order separated by 1 s (Fig. 1) and were instructed to report which appeared to be larger by means of left/right mouse buttons—a categorization separate from the stimulus dimension concerning size, which reduced the potential for response bias. The test circle had a fixed diameter and was accompanied by a time-staggered tone. The probe circle was accompanied by a simultaneous tone, and its diameter was varied according to a two-interval, forced-choice QUEST adaptive paradigm (Watson and Pelli 1983) for each audiovisual delay (Fig. 2). Each of the two experiments took approximately 30 min ($\approx$300 trials). In a piloting phase of the experiment, participants reported that all audiovisual stimuli appeared simultaneous in terms of the onset of their audio and visual components.

Data analysis

Points of subjective equality (PSEs) were derived by fitting a sigmoid to the set of sizes chosen by the QUEST over 35 trials for each sound onset asynchrony (SOA). For each participant at each asynchrony, the PSE was normalized by subtracting the PSE for the simultaneous condition (test circle with synchronous tone) from the PSEs obtained in each of the asynchronous tone conditions.

Results

Figure 2 depicts the change in PSE across the different audiovisual asynchronies relative to the simultaneous condition. When sound onset lagged visual onset (positive asynchronies, data from experiment one, filled circles), a planned polynomial contrast revealed a significant linear trend in the predicted direction ($t = 2$, $P = 0.03$, $df = 1$). Specifically this analysis revealed larger perceived sizes for longer audiovisual delays with an increase in PSE between the 20 and 80 ms SOAs ($t = 2.74$, $P = 0.007$, $df = 15$, Bonferroni corrected). A significant Pearson correlation between SOA (stimulus onset asynchrony) and perceived size was also determined ($r = 0.82$, $P = 0.02$, $n = 7$) within this SOA range (20–80 ms). The effect of sound delay decreased when larger asynchronies were tested in experiment two. In particular, testing the PSEs obtained over an extended range of SOAs between $\pm$128 ms, we observed an increase in the perceived size of the test and probe circles that peaked at an asynchrony of +64 ms ($t = 4.19$, $P = 0.003$, $df = 8$). No significant change in perceived size was found when sound onset lagged by 128 ms or when sound onset led visual onset (negative SOAs).

Discussion

The perceived size of the test circle, as indicated by the size of the probe that was matched to it, reliably increased with the addition of a delayed sound up to an asynchrony of about 80 ms. This increase was replicated when a larger range of asynchronies was employed. No effect was observed for negative SOAs (sound leading) over the range tested. Despite the different ranges and interval sizes tested in each experiment, the pattern of change from baseline did not differ

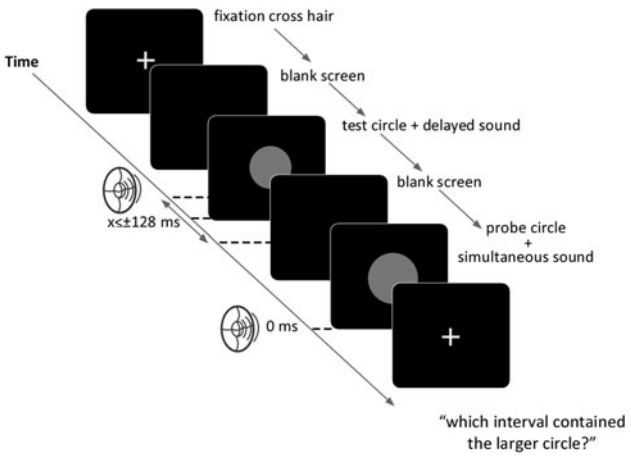

Fig. 1 Trial sequence. Each trial began with a fixation crosshair (1.5 s) followed by a 1 s pause before the presentation of the first stimulus interval. This example illustrates a sequence for which a test circle (with variable tone delay) was presented in the first interval and the probe circle (with simultaneous tone) in the second. The QUEST adaptive procedure was used to vary the size of the probe circle to determine when it appeared to match the size of the test circle

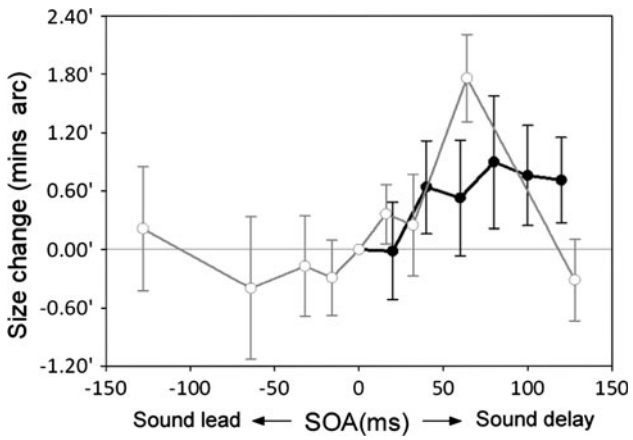

Fig. 2 Mean judgments of the change in size of the probe circle relative to the simultaneous sound condition plotted as a function of audiovisual asynchrony. *Closed circles* and *solid lines* represent mean size change of participants tested with positive asynchronies only (experiment one, sound lagging). *Open circles* and *dashed lines* represent the mean size change from experiment two using a larger range of asynchronies (including sound leading). *Error bars* are standard error of the mean size change

qualitatively between the two data sets within the overlapping range of SOAs. That no effect was found for negative SOAs rules out the influence of a simple response bias toward responding "larger" for asynchronous events. These data show "therefore" that the perceived visual size of an audiovisual event depends partially on the relative timing of the presentation of the audio and visual components.

The influence of audiovisual asynchrony on perceived size is novel. A non-specific enhancement of visual properties by an accompanying sound (see Shams and Kim 2010 for a review) would be largest when the audio and visual components were approximately synchronous (in line with the "temporal constraint" on sensory integration, Meredith et al. 1987). Any such enhancement effect would be expected to fall off symmetrically as the components became more asynchronous eventually disappearing. The present effect was instead highly asymmetrical and did not peak when the components were synchronous. Specifically, we demonstrated an increase in the perceived size of a visual stimulus as the onset of the sound was increasingly delayed over the range between 0 and 120 ms. Thus, rather than demonstrating a non-specific enhancement effect, our data suggest that subjects are sensitive to the difference in arrival times of light and sound when the sound arrives after a delay. Such arrival time differences arise in the environment because of the slow speed of sound relative to light resulting in sound arriving after a delay that depends on the distance of the event from the observer. Why might the apparent size of a circle increase when presented with a slightly delayed sound?

We contend that the perceived increase may be related in part to an induced change in the perceived distance of the audiovisual event that could theoretically be deduced from the relative timing of its auditory and visual components. For a given retinal image, the perceived size of the object giving rise to it varies with perceived distance: a phenomenon commonly referred to as size-distance invariance (Schlosberg 1950; Gilinsky 1951; Kilpatrick and Ittelson 1953; Gilinsky 1955; Epstein et al. 1961; Kaufman et al. 2006; Kaufman et al. 2007). Thus, our data are compatible with test circles presented with a slightly delayed sound appearing more distant (and therefore being perceived as larger) than a circle presented with a simultaneous sound (see Fig. 3). The lack of similar effects in perceived size with leading sounds may be due to the fact that sound cannot normally arrive at an observer before the visual component arising from an audiovisual event. Such stimuli are ecologically invalid. This line of reasoning is supported by evidence showing that audiovisual neurons exhibit increased responses to asynchronous stimuli are typically more sensitive to sound delays than to leads (Meredith et al. 1987).

The shift in perceived distance required to explain the size changes we report is considerably smaller (peaking

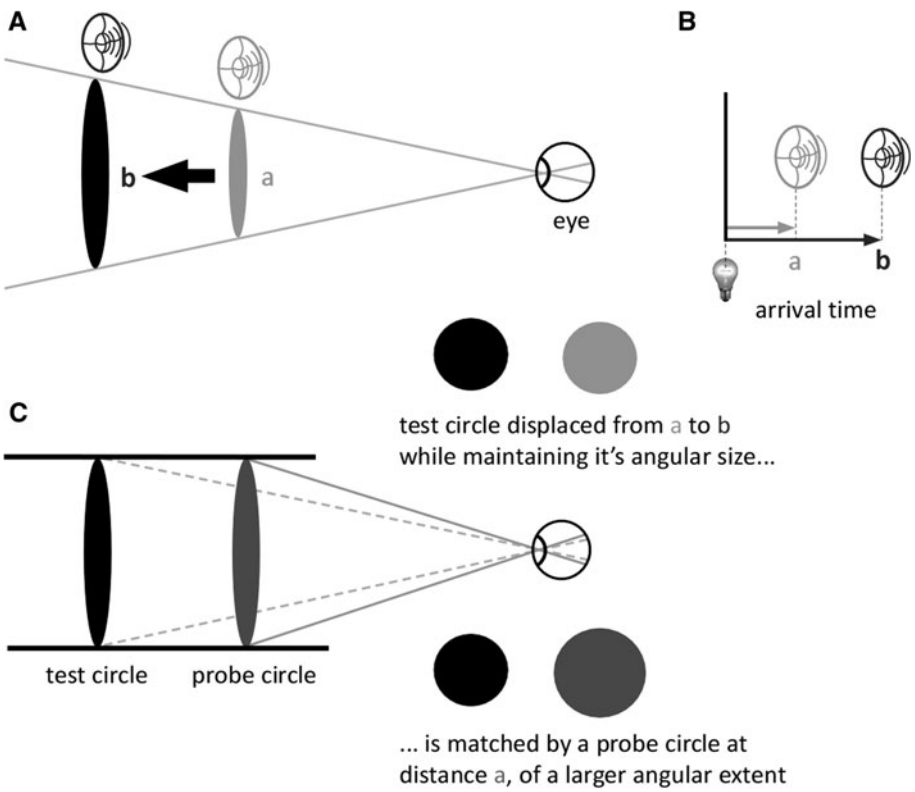

Fig. 3 Interpreting the data shown in Fig. 2 using size-distance invariance. (**a**) A particular retinal image could be created by objects of various sizes depending on their distance from the observer. (**b**) The delay between the time of arrival of the visual and auditory components of an event increases linearly with distance. (**c**) A change in perceived target distance causes an object to change its perceived size

around 6–12 cm) than the distances that should correspond to the audiovisual delays employed if the speed of sound (340 m/s) were fully accounted for (for example, an audiovisual asynchrony of 64 ms corresponds to a viewing distance of about 19 m). It is possible that the magnitude of this cross-modal contribution to perceived distance might be greater for real-world events with more complex spatiotemporal and spectral cues or at larger distances than the one used here. In our experiments, other possible distance cues that our participants could be using, such as accommodation cues and knowledge about the size of rooms, may have conflicted with the audiovisual delay cue and reduced the strength of the effect seen here. Although our effects were small, we have demonstrated that cross-modal delay can alter distance/size judgments in a manner consistent with the audiovisual delay information being available to the observers. Thus, audiovisual delay joins other significant yet subsidiary cues (e.g., retinal blur; Vishwanath and Blaser 2010) in contributing to solving the formidable problem of determining how far away something is.

Specifically, the size increase induced by sound delay was found only when sound was delayed by between 0 and 120 ms: for longer delays the effect diminished. We interpret this restricted range as resulting from the restricted temporal integration window for audiovisual fusion (reviewed in Harris et al. 2010). That is, successful extraction of distance from audiovisual delay can only occur over the range of delays where the auditory and visual components are likely to be fused into a single event. As sound delay increases beyond this range, the likelihood of audiovisual integration decreases and therefore the relative timing of these separate events becomes irrelevant in providing additional information about either one. The temporal envelope observed here is in agreement with the window of visual influence of sound on synesthetically matched tones and visual circles found by Parise and Spence (2009). Their study showed a wider window of integration with a median just noticeable difference in audiovisual temporal order judgments of about 85 ms for congruently paired circle size/tone frequency combinations relative to incongruent combinations, reflecting an enhancement in audiovisual interaction.

The trade off between cross-modal delay signaling distance and the limitation imposed by the temporal window of integration is reflected in the initial linear increase in PSE with positive asynchrony and then the drop off at longer delays. Using the data from the original experimental setup, we model the trade off as a weighted average between a linear effect of asynchrony with a slope measured in secs arc/ms (where a slope of 123 s arc/ms represents a perfect extraction of distance information) and no effect (with a slope of 0 s arc/ms). The relative weights assigned to these two strategies (linear: null) vary from 100:0 at 0 delay through to 0:100 at some longer delay. There are three free parameters in this model: the linear slope, the rate of fall off (the weights are modeled, arbitrarily, as varying sigmoidally—see Fig. 4), and the delay at which equal weight is given to both: the 50:50 point. The variation in weights that best describes the data is shown graphically in Fig. 4a. The best fit to our observed data (shown in Fig. 4b) was obtained with a linear effect of 1 s arc/ms (representing 0.8% of perfect distance extraction) and equal weighting occurring at 120 ms. The model predicts a drop off after 100–110 ms, which is the result that was independently observed in the second experiment.

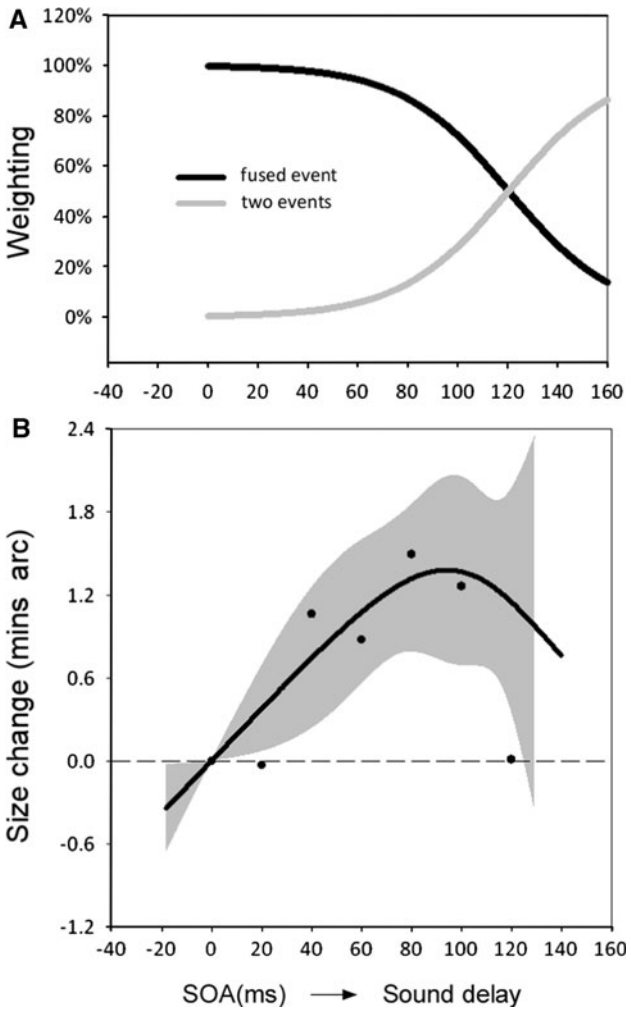

Fig. 4 Model of our data as a weighted sum between a linear effect when the auditory and visual components were fused and a null effect when they were beyond the fusion range (**a**). Weights were varied sigmoidally where the *black line* is the relative weighting applied to the linear function and the *gray line* is the relative weighting applied to the null effect. The best fit was found when x_0 (equal weighting of the two functions) was at 120.7 ± 14 ms, and the slopes of the sigmoids were 11 ± 9 ms, and the linear effect (plateau of the sigmoid) was 0.95 ± 0.28 s arc/ms ($r^2 = 0.84$) (**b**). 95% confidence limits are shown suggesting delays of less than 120 ms produced a significant size increase

Other factors such as spatial proximity (i.e., the sound and light need to be close enough in space to be regarded as originating from a common source), temporal resolution (the brain needs to be able to resolve the time differences involved), and potentially conflicting cues to distance within our experimental design (e.g., the fact that we used sounds of constant intensity; participants could use accommodation cues to the actual distance of the screen) may also contribute to the structure and magnitude of the function we report. Nevertheless, the finding that perceived size is altered by audiovisual asynchrony demonstrates a striking sensitivity to audiovisual time-of-arrival differences and may reveal a potentially important new cue to perceived distance: size/distance computation can involve cross-modal interaction beyond the use of unisensory information alone.

Acknowledgments LRH is supported by the Natural Sciences and Engineering Research Council of Canada. SS-F and PJ are supported by the *Spanish Ministry of Science and Innovation* (PSI2010-15426 and Consolider INGENIO CSD2007-00012), *Comissionat per a Universitats i Recerca del DIUE-Generalitat de Catalunya* (SRG2009-092), and the European Research Council (StG-2010 263145).

References

Brainard DH (1997) The psychophysics toolbox. Spat Vis 10:433–436

Cappe C, Thut G, Romei V, Murray MM (2009) Selective integration of auditory-visual looming cues by humans. Neuropsychologia 47:1045–1052

Epstein W, Park J, Casey A (1961) The current status of the size-distance hypotheses. Psychol Bull 58:491–514

Gallace A, Spence C (2006) Multisensory synesthetic interactions in the speeded classification of visual size. Percept Psychophys 68:1191–1203

Gilinsky AS (1951) Perceived size and distance in visual space. Psychol Rev 58:460–482

Gilinsky AS (1955) The effect of attitude upon the perception of size. Am J Psychol 68:173–192

Gregory RL (1997) Eye and brain. Princeton University Press, New Jersey

Harris LR, Harrar V, Jaekl P, Kopinska A (2010) Mechanisms of simultaneity constancy. In: Nijhawan R, Khurana B (eds) Space and time in perception and action. Cambridge University Press, Cambridge

Hershenson M (ed) (1989) The moon illusion. Lawrence Erlbaum & Associates, New Jersey

Kaufman L, Kaufman JH, Noble R, Edlund S, Bai S, King T (2006) Perceptual distance and the constancy of size and stereoscopic depth. Spat Vis 19:439–457

Kaufman L, Vassiliades V, Noble R, Alexander R, Kaufman J, Edlund S (2007) Perceptual distance and the moon illusion. Spat Vis 20:155–175

Kilpatrick FP, Ittelson WH (1953) The size-distance invariance hypothesis. Psych Rev 60:223–231

King AJ (2005) Multisensory integration: strategies for synchronization. Curr Biol 15:R339–R341

Kitagawa N, Ichihara S (2002) Hearing visual motion in depth. Nature 416:172–174

Kopinska A, Harris LR (2004) Simultaneity constancy. Perception 33:1049–1060

Meredith MA, Nemitz JW, Stein BE (1987) Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. J Neurosci 7:3215–3229

Parise CV, Spence C (2009) When birds of a feather flock together: synesthetic correspondences modulate audiovisual integration in non-synesthetes. PLoS One 4:e5664

Ross HE, Plug C (2002) The mystery of the moon illusion: exploring size perception. Oxford University Press, Oxford

Schlosberg H (1950) A note on depth perception, size constancy, and related topics. Psychol Rev 57:314–317

Shams L, Kim R (2010) Crossmodal influences on visual perception. Phys Life Rev 7:269–284

Spence C, Squire S (2003) Multisensory integration: maintaining the perception of synchrony. Curr Bio 13:R519–R521

Valjamae A, Soto-Faraco S (2008) Filling-in visual motion with sounds. Acta Psychol (Amst) 129:249–254

Vishwanath D, Blaser E (2010) Retinal blur and the perception of egocentric distance. J Vis 10:1–16

Watson AB, Pelli DG (1983) QUEST: a Bayesian adaptive psychometric method. Percept Psychophys 33:113–120